

Emoticon based Sentiment Analysis using Parallel Analytics on Hadoop

V. Jude Nirmal¹ and D. I. George Amalarethnam²

¹Department of IT, St. Joseph's College (Autonomous), College Road, Tiruchirappalli - 620002, Tamil Nadu, India; nirmaljude6@gmail.com

²Jamal Mohamed College, 7, Race Course Road, Khajanagar, Tiruchirappalli - 620020, Tamil Nadu, India; di_george@ymail.com

Abstract

Objectives: The major objective of this approach is to provide a sentiment analysis architecture that can operate on streaming big data to provide effective results at tolerable time limits. **Method/Analysis:** An effective mechanism for analyzing the social networking messages to identify the sentiment levels has been proposed. This is a generic model that can be used for product or organization specific analysis. Further, this method also considers emoticons, which form the integral part of any expressed emotion. The entire process is carried out in Hadoop Architecture using the MapReduce paradigm. **Findings:** Experiments have been conducted on a Hadoop cluster. Inputs were passed from a client node connected to the cluster. Map Reduce programs were executed in six phases, each phase performing a single task in map and reduce phases. The ROC plot exhibits excellent accuracies with most of the points being clustered in the top left region, some even approaching 100% effectiveness. Even the PR plots exhibit similar efficiency scenario with high positive retrieval rates. Incorporating the emoticons plays a major role in increasing the efficiency of this approach. **Novelty/Improvement:** This approach uses Hadoop based implementations, involving Map and Reduce operations. Using this approach provides data scalability and improves the efficiency of the results in acceptable time limits.

Keywords: Emoticons, Hadoop, Map Reduce, Polarity Identification, Sentiment Analysis, Social Networking Data Processing

1. Introduction

Sentiment analysis has become one of the major necessities of today's business industry. Business, both online and offline require major inputs from consumers in terms of feedbacks. Not many consumers provide direct feedback. In order to formulate effective business strategies, consumer requirements/feedbacks are needed. The advent of social networks and mass adoption of them has presented a major opportunity for these organizations to leverage the interests of the user and convert it to business strategies.

Though user opinions are reflected in the messages posted in social networking sites, these messages are textual and do not exhibit direct correspondence to the products that are dealt with by the organizations. Correlation between opinions exhibited by public and products manufactured by the companies is the missing

link in this scenario¹. The strategies proposed for constructing this missing link is called text mining. In order to be specific, the methods concentrating on mining social networking data tends to identify the major player or contributor in the sentence, along with identifying the polarity of the sentence, which helps identify the sentiment levels of the sentence.

Several classification based methods for sentiment identification have been proposed. A classifier ensemble based sentiment identification method², which considers the query term and classifies the tweets as positive or negative. This method has its major concern in product or organization based analysis. It uses Multinomial Naive Bayes, SVM, Random Forest, and Logistic Regression for classification. The advantage of this approach is that it provides best results due to the combination of several algorithms. The downside of this approach is that it tends to increase the processing overhead in

* Author for correspondence

manifolds. A similar feature based learning model³ provides effective sentiment analysis. A multi-lingual based sentiment analysis technique⁴, that has its major concerns in operating with multiple language support for effective results. This paper is based on two languages, English and Spanish. This method analyzes the hybrid features and explore multi-lingual data, and it boasts of improved accuracies. A context based semantic analysis of twitter data⁵, that operates on both entity level and tweet level, providing enhance accuracy. Similar context based sentiment analysis methods were presented in^{6, 7}. This method proposes to exhibit high robustness to noise, by eliminating them, hence providing better analysis. A similar twitter based sentiment analysis method was presented in⁸. A sentiment based opinion generation technique was proposed in⁹. Behavior analysis methods^{10,11} are also prevalent, that takes the sentiment analysis to next level. Emotion tracking and time series emotion tracking provides an enhanced way to track social/product sentiments. A time series based emotion tracking¹², that models emotions on the basis of time. Several such techniques tracking emotions on the basis of time has been proposed¹³⁻¹⁵. User specific emotion analysis also exist in literature that tracks the emotions of a single user¹⁶. Irony in text proves to be the most complex structure to be evaluated. Some effective techniques working on detecting irony includes¹⁷⁻²⁰.

2. System Architecture

The process of sentiment extraction from documents plays a vital role in the business analytics. Texts from social networking/ microblogging sites like Twitter gain much more prominence due to the fact that they reflect the current and happening scenarios. The architecture depicting sentiment identification in documents is presented in Figure 1.

The sequence of document processing for sentiment analysis is depicted in Figure 1. The input tweets/documents are passed through two processing scenarios. The first part is to identify the sentiment from text and the second part is identifying the polarity from the emoticons available in the text. The tweets are duplicated and passed to both the scenarios. The first process eliminates all the symbols present in the text and performs stemming, normalization and lemmatization to find the polarity, while the second process extracts the emoticons to identify the polarity. Both the polarity values are combined to

identify the aggregated polarity of the document under scrutiny.

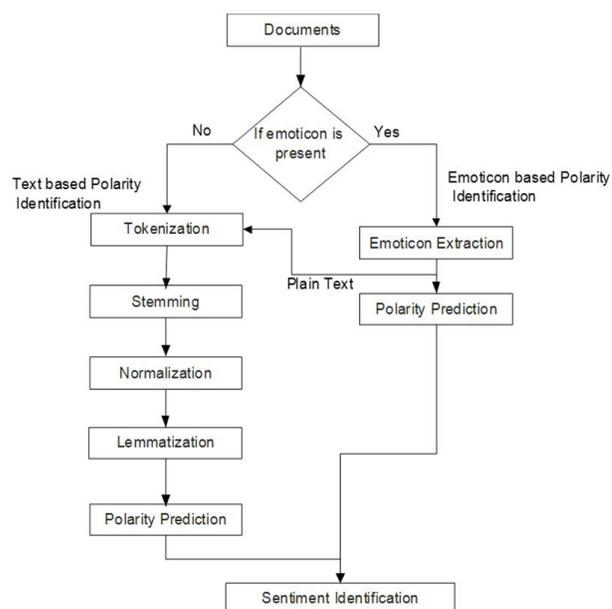


Figure 1. Sentiment Identification Architecture.

3. Proposed Approach

Sentiment analysis of user comments or feedbacks have played a vital role in successful marketing of products. Though the initial analysis were conducted on the text retrieved from organization specific web sites, it was later identified that social networking sites provide much better and unbiased results. Current business analysis are all conducted using social networking data. The social networking data tends to be streaming data, generated at large. This necessitates the use of Big Data techniques. The current method uses HDFS to store the data and MapReduce Paradigm is used to process the data returning the sentiment indicators. In the current social networking scenario, data tends to be not just text but also a group of symbols known as emoticons^{21,22}. The emoticons also play a vital role in identifying the polarity of the current document under scrutiny²³. Hence our method uses both text and emoticons as the base indicators to identify the sentiment value.

The process of sentiment identification is carried out in four phases; document analysis and segregation, document processing, emoticon processing and sentiment analysis.

3.1 Document Analysis and Segregation

The first phase of processing is the document analysis. Base documents from Twitter are passed on to the system. The system performs two major processing; document processing and emoticon processing. Hence it should decide if the current text contains emoticons in order to branch for both the process. Else a document based processing alone would suffice. The decision making process is performed here. If the text contains emoticons, it is duplicated and is passed to both document processing and emoticon processing else it is passed to document based processing alone.

3.2 Document Processing

The document processing of tweets eliminates the symbols and processes the other constituents of the tweet to identify the polarity of the tweet under analysis.

3.2.1 Tokenization

Tokenization is the process of dividing a string into stream of text; words, symbols, phrases and other meaningful elements. The entire contents of the tweet are taken and several filters are applied to them sequentially to eliminate special characters, symbols and spaces from the text. The output returned from the tokenizer is usually in the form of a vector, containing all the tokens corresponding to the tweet under analysis. Composition of these output tokens are limited to alphabets alone.

3.2.2 Stemming

Stemming is the process of eliminating affixes (prefix and suffix) from a word, to obtain the seed word. It is basically used to reduce the inflection in the words. The most popular and the standard algorithm for stemming is the Porter Stemming algorithm¹⁸, proposed by Martin Porter. Several versions of this algorithms also exist in literature. This paper uses the Porter Stemmer 2, the modified version of the original porter algorithm to perform stemming.

3.2.3 Normalization/ Lemmatization

Normalization is the process of converting a text to a canonical format, such that uniformity can be achieved in the corpus. This makes sure the text that is stored is consistent and is reliable in order to perform higher level operations. Lemmatization is the process of

grouping together several inflected forms of the same word into a single word. Both the processes are similar in context, hence they are performed in a single phase. Normalization and lemmatization almost similar, hence they are combined and performed as a single process. WordNet 3.0 database is used as the reference for the process of normalization and lemmatization.

3.2.4 Polarity Prediction

After the completion of tokenization, stemming, normalization and lemmatization, the returned tokens are of the root/ base form of the words that were used in the actual text. Polarity of terms is obtained from WordNet 3.0, which is a human curated dataset for polarity analysis. This dataset is the base for all operations, hence to parallelize the entire operation, the data is stored in the distributed cache in Hadoop Distributed File System(HDFS). This operation distributes the polarity data to all the data nodes in the Hadoop system, hence reducing the need for communications and references.

3.3 Emoticon Analysis

This process is activated on identifying emoticons, in the input tweet. The basic operation in this section is to identify the emoticons present in the text. Emoticons are groups of symbols expressing an emotion.

Table 1. Emoticons and their polarity

Emoticon	Polarity	Emoticon	Polarity
:-)	1	:'(-1
:)	1	:'-)	1
:D	1	:')	1
:]	1	:-O	1
:>	1	:O	1
=]	1	:-o	1
:-D	1	:o	1
:-))	1	:*	1
:-(-1	;-)	1
:(-1	;))	1
:-<	-1	;-]	1
<:	-1	;]	1
:[-1	;D	1
{	-1	:P	1
;(-1	:-p	1
:-	-1	:p	1
@	-1	:-/	-1
:'-)	-1	:/	-1
<3	1	:\	-1
</3	-1		

The current process uses 39 most commonly used emoticons to identify the polarity. Table 1 shows the list of emoticons used for analysis. The advantage of using this approach is that the emoticons are not directly used for reference. Instead, they are used for training the classifier and the classifier rules are used to identify the polarity, depending on the emoticon being encountered in the tweet.

Decision Table is used as the classifier for training the system. Decision table is a precise way to model complex set rules and their corresponding actions. A decision table is usually constructed using conditional clauses defining the conditions followed by the action entries. The decision table is constructed by using the 10-fold-cross-validation technique. Figure 2 shows a snapshot of the rule that has been built by the decision tree.

3.4 Sentiment Identification

The sentiment identification phase is the final phase of sentiment analysis. The input tweets are passed to the text based polarity identification phase and if emoticons are present, they are extracted and decision rules are used to identify the polarity of the emoticon. This phase aggregates the polarity values obtained in the prior phases to provide the aggregated polarity of the text. This aggregated polarity identifies the sentiment of the current tweet. Since specific values are used to identify the sentiment of the text, the resultant value not only depicts the polarity, but also the intensity of the polarity. This intensity defines the impact the tweet makes on a user. Eg. A tweet with an intensity of -0.02 has low negative impact, while a tweet with intensity -0.78 has a high negativity associated with it.

Several methods tend to identify the polarity of the text alone, without identifying the intensity of the polarity. The proposed approach provides intensity based calculations, using which the polarity along with the sentiment intensity can be obtained.

4. Results and Discussion

Experiments were conducted to analyze and identify the accuracy of the proposed method. Inputs were passed from a client node connected to the cluster. Map Reduce programs were executed in six phases, each phase performing a single task in map and reduce phases. Execution plan of the current process is presented in Figure 3. Experiments were conducted using the STS Gold Sentiment Corpus²⁴ to identify the efficiency of our method. Experiments were carried out on a 12 Node Hadoop Cluster, running Hadoop 1. The cluster constitutes 10 Data Nodes, a Name Node and a Secondary Name Node.

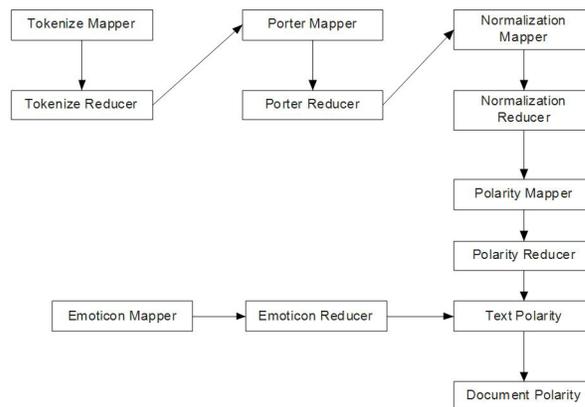


Figure 3. Sentiment Identification Execution Plan.

```

if data[0]==':' and data[1]=='-' and data[2]=='D':prediction=1if
data[0]==':' and data[1]=='-' and data[2]=='(':prediction=-1if data
[0]==':' and data[1]=='p':prediction=1if data[0]==';' and data[1]
=='D':prediction=1if data[0]==':' and data[1]=='(':prediction=-1if
data[0]==':' and data[1]=='o':prediction=1if data[0]==':' and data
[1]=='D':prediction=1if data[0]==':' and data[1]=='*':prediction=1if
data[0]==':' and data[1]=='@':prediction=-1if data[0]==':' and data
[1]=='-' and data[2]=='o':prediction=1if data[0]=='=' and data[1]
=='j':prediction=1if data[0]==';' and data[1]=='-' and data[2]
=='j':prediction=1if data[0]==':' and data[1]=='':prediction=1if
data[0]==':' and data[1]=='-' and data[2]=='/':prediction=-1if data
[0]=='<' and data[1]=='3':prediction=1if data[0]==':' and data[1]
=='-' and data[2]=='o':prediction=1if data[0]==';' and data[1]
=='(':prediction=-1if data[0]==':' and data[1]=='[':prediction=-1if
data[0]==':' and data[1]=='-' and data[2]==')' and data[3] ...
  
```

Figure 2. Decision Rules (Sample).

Table 2. Confusion Matrix and Calculated Data

TP	FP	TN	FN	FPR	TPR	Recall	Precision	F1	Accuracy
0	0	1	0	0	0	0	0	0	1
0	0	2	0	0	0	0	0	0	1
0	0	3	0	0	0	0	0	0	1
0	1	3	0	0.25	0	0	0	0	0.75
0	2	3	0	0.4	0	0	0	0	0.6
0	2	4	0	0.333333	0	0	0	0	0.666667
0	3	4	0	0.428571	0	0	0	0	0.571429
0	3	5	0	0.375	0	0	0	0	0.625
0	3	6	0	0.333333	0	0	0	0	0.666667
.....									
241	453	928	387	0.328023	0.383758	0.383758	0.347262	0.364599	0.581882
241	453	929	387	0.327786	0.383758	0.383758	0.347262	0.364599	0.58209
241	453	930	387	0.327549	0.383758	0.383758	0.347262	0.364599	0.582297
241	453	931	387	0.327312	0.383758	0.383758	0.347262	0.364599	0.582505
241	453	932	387	0.327076	0.383758	0.383758	0.347262	0.364599	0.582712
241	453	933	387	0.32684	0.383758	0.383758	0.347262	0.364599	0.58292
241	453	934	387	0.326604	0.383758	0.383758	0.347262	0.364599	0.583127
241	453	935	387	0.326369	0.383758	0.383758	0.347262	0.364599	0.583333
241	453	936	387	0.326134	0.383758	0.383758	0.347262	0.364599	0.58354
241	453	937	387	0.325899	0.383758	0.383758	0.347262	0.364599	0.583746
241	453	938	387	0.325665	0.383758	0.383758	0.347262	0.364599	0.583952
241	453	939	387	0.325431	0.383758	0.383758	0.347262	0.364599	0.584158
241	454	939	387	0.325915	0.383758	0.383758	0.346763	0.364324	0.583869

The STS Gold Sentiment Corpus is a human annotated dataset containing 2034 tweets. Predictions obtained from and their corresponding confusion matrices along with the calculated values are shown in Table 2.

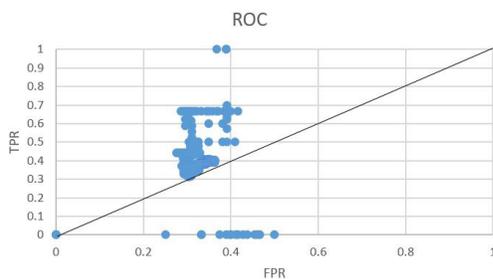


Figure 4. ROC Curve.

Figure 4 shows the Receiver Operating Characteristic (ROC) curve plotted using the STC gold corpus data. It could be observed that most of the plotted coordinates are grouped in the top left corner of the graph, above the diagonal line. Some points even approach the top left (0,1) coordinate, which proves that the current method exhibits effective results.

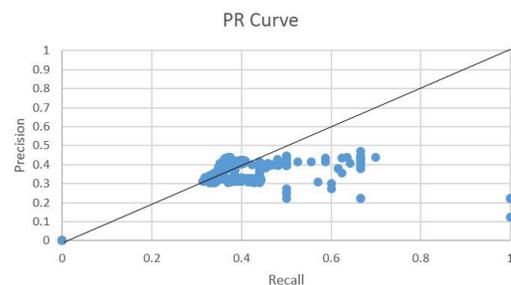


Figure 5. PR Curve.

Figure 5 shows the (Precision Recall Curve) PR curve. It represents the accuracy of the results obtained and the accuracy of the selection mechanism. This method shows that most of the points are concentrated below the diagonal, which shows that the results retrieved from the solution set are accurate with least false positives.

Figure 6 presents the accuracy levels obtained by the process. High accuracy levels could be observed.

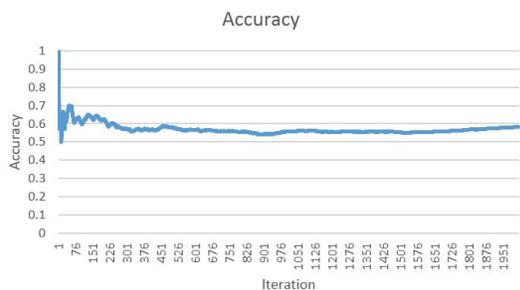


Figure 6. Accuracy.

5. Conclusion

This paper presents effective mechanisms to perform sentiment analysis on streaming data. Since social networking has become a huge storehouse of information, it becomes difficult to perform processing on it. Our proposal uses Hadoop to perform analysis on tweets to identify their polarities. Hadoop is a technique used to perform processing on Big Data, since our application qualifies as Big Data, the need for Hadoop is justified. Rather than utilizing only the text for polarity identification, the emoticons are also taken as meaningful entities to identify the polarity. This provides better accuracy in defining the polarity. Both the polarity vectors are aggregated and the final sentiment level is identified. Further, our research can be extended by using online processing techniques such as Spark to enable processing of real time streaming data and provide faster results.

6. References

1. Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*. 2015 Nov 30; 89:14-46.
2. da Silva NF, Hruschka ER, Hruschka ER. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*. 2014 Oct 31; 66:170-9.
3. Baecchi C, Uricchio T, Bertini M, Del Bimbo A. A multi-modal feature learning approach for sentiment analysis of social network multimedia. *Multimedia Tools and Applications*. 2015; p. 1-9.
4. Balahur A, Perea-Ortega JM. Sentiment analysis system adaptation for multilingual processing: The case of tweets. *Information Processing & Management*. 2015 Jul 31; 51(4):547-56.
5. Saif H, He Y, Fernandez M, Alani H. Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management*. 2016 Jan 31; 52(1):5-19.
6. Katz G, Ofek N, Shapira B. ConSent. *Knowledge-Based Systems*. 2015 Aug 1; 84(C):162-78.
7. Korenek P, Simko M. Sentiment analysis on microblog utilizing appraisal theory. *World Wide Web*. 2014 Jul 1; 17(4):847-67.
8. Ruba KV, Venkatesan D. Building a Custom Sentiment Analysis Tool based on an Ontology for Twitter Posts. *Indian Journal of Science and Technology*. 2015 Jul; 8(13):1-9.
9. Zol S, Mulay P. Analyzing Sentiments for Generating Opinions (ASGO)-A New Approach. *Indian Journal of Science and Technology*. 2015 Feb; 8(S4):1-6.
10. Mahajan C, Mulay P. E3: Effective Emoticon Extractor for Behavior Analysis from Social Media. *Procedia Computer Science*. 2015 Dec 31; 50:610-6.
11. Litvinova TA, Seredin PV, Litvinova OA. Using Part-of-Speech Sequences Frequencies in a Text to Predict Author Personality: a Corpus Study. *Indian Journal of Science and Technology*. 2015 May; 8(S9):1-5.
12. Zhu C, Zhu H, Ge Y, Chen E, Liu Q, Xu T, Xiong H. Tracking the evolution of social emotions with topic models. *Knowledge and Information Systems*. 2015 Jul 28; p. 1-28.
13. Chaumartin FR. UPAR7: A knowledge-based system for headline sentiment tagging. *Association for Computational Linguistics: Proceedings of the 4th International Workshop on Semantic Evaluations*, PA. 2007 Jun 23; p. 422-25.
14. Kozareva Z, Navarro B, Vazquez S, Montoyo A. UA-ZBSA: a headline emotion classification through web information. *Association for Computational Linguistics: Proceedings of the 4th International Workshop on Semantic Evaluations*, Spain. 2007 Jun 23; p. 334-37.
15. Lin KH, Yang C, Chen HH. What emotions do news articles trigger in their readers? *ACM: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval NY*. 2007 Jul 23; p. 733-34.
16. Bao S, Xu S, Zhang L, Yan R, Su Z, Han D, Yu Y. Joint emotion-topic modeling for social affective text mining. *ICDM'09, Ninth IEEE International Conference on Data Mining, 2009*. IEEE, 2009 Dec 6; p. 699-704.
17. Reyes A, Rosso P, Veale T. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*. 2013 Mar 1; 47(1):239-68.
18. Kreuz R. Using figurative language to increase advertising effectiveness. University of Memphis, Memphis, TN: Office of Naval Research Military Personnel Research Science Workshop. 2001 Jun 4.
19. Kumon-Nakamura S, Glucksberg S, Brown M. How about another piece of pie: The allusional pretense theory of discourse irony. London: Taylor and Francis Group: Gibbs R, Colston H (Eds.). *Irony in language and thought*. 2007; p. 57-96.
20. Lucariello J. Situational irony: A concept of events gone away. *Irony in language and thought*. 2007; p. 467-98.

21. Shukla A, Chaudhary BD. A study of usage of symbols and opinionated words in annotation for modeling literature survey experiences. *Education and Information Technologies*. 2015 Mar 1; 20(1):91-111.
22. Xiong X, Zhou G, Huang Y, Chen H, Xu K. Dynamic evolution of collective emotions in social networks: a case study of Sinaweibo. *Science China Information Sciences*. 2013 Jul 1; 56(7):1-8.
23. Feng S, Song K, Wang D, Yu G. A word-emoticon mutual reinforcement ranking model for building sentiment lexicon from massive collection of microblogs. *World Wide Web*. 2015 Jul 1; 18(4):949-67.
24. Saif H, Fernandez M, He Y, Alani H. Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. 2013; p. 1-13.